Change-point Analysis

This course will cover classic works and recent advances on change-point analysis, both *offline* change-point analysis (also referred to as *segmentation*), which divides a completely observed sequence into homogeneous temporal or spatial segments, and *online/sequential* change-point analysis, which on-the-fly detects changes in sequentially observed data. Change-point detection is a classic problem and was extensively studied for the univariate case. However, with advances in technologies, the collection of massive data becomes feasible and new challenges arise as the dimension of the observations in the sequence becomes higher or the observations are even non-Euclidean. This triggers a recent wave of change-point method developments for modern data.

This course will first prepare students with standard techniques for change-point problems, which are also useful for studying modern change-point problems. This course will then explore recent developments on change-point detection for multivariate data and for non-Euclidean data. The detailed contents are listed below.

Part I: **Traditional change-point analysis for univariate data.**

In this part, the main concepts of offline and online change-point analysis will be covered. It will first start with the easiest scenario that the observations follow the Gaussian distribution with known variance and the change is a mean shift. Classical procedures will be covered as well as major theoretical results. Then, the scenario will be relaxed to Gaussian distribution with unknown variance and then to other distributions. Some recent applications on genomic data will be covered.

Part II: **Change-point analysis for multivariate data.**

This part will consist two sub-topics: (i) The multiple sequences are independent, and (ii) The multiple sequences are not independent. We focus on parametric methods in this part. For scenario (i), current methods can deal with quite high dimensions and we will cover ways of effectively integrating information from the multiple sequences. For scenario (ii), current methods could deal with low-dimensional data. Some successful examples for both scenarios for recent studies in genomic data and multiple sensor problems will be covered.

Part III: **Recent advances on change-point analysis for high-dimensional data/non-Euclidean data.**

In this part, non-parametric methods will be covered, namely, methods based on similarity of the observations, which can be applied to data in arbitrary dimension and to non-Euclidean data. The main techniques, which are quite different from parametric approaches, will be covered. Applications on modern data, such as network data, will be discussed.